Deep Learning Project on Image Classification

StudentA
Shien-Ming Wu School of Intelligent Engineering
South China University of Technology
Canton, China
****@***.***

StudentB (repository maintainer)
School of Computer Science & Engineering
South China University of Technology
Canton, China
****@***.***

StudentC
Shien-Ming Wu School of Intelligent Engineering
South China University of Technology
Canton, China
****@***.***

Abstract—**In this project, we used tools like Labelme and Roboflow to annotate images. By combining ResNet50 with a Vision Transformer (ViT), we managed to harness the strengths of both models, effectively offsetting their individual weaknesses in dealing with noise and occlusion. The final model reached an impressive 90% accuracy.**

Keywords—image classification, deep learning, computer vision, hybrid neural networks, convolutional neural networks, vision transformer, model ensemble

I. Introduction

With the high-speed developing of computer vision technology, deep learning achieved remarkable success in image classification. As one of the fundamental tasks of computer vision, image classification is widely used in many domains such as object recognition, intelligent surveillance, autonomous driving… This project aims to classify six objects related to South China University of Technology Guangzhou International Campus by using deep learning techniques. Those objects, including bell tower, library, logo of Shien-Ming Wu School of Intelligent Engineering, Liyujun mascot, Mingcheng mascot and Junde mascot, are not only form the important part of campus culture but also possess unique visual characteristics and recognition value.

The main goal of this project is to build an efficient and accurate deep learning model to classify the six objects automatically. Therefore, we completed the following core tasks: curate a high-quality and diverse dataset by taking pictures and using photo downloaded from the Internet. Then, we design and train a model with strong generalization capabilities. Finally, we evaluate the model's performance including accuracy, operational efficiency and robustness. Through the practice of this project, we aim to gain an in-depth understanding of the application of deep learning in image classification and actively explore innovative methods and techniques.

II. Dataset Construction and Preprocessing

High quality dataset is critical of strong model performance. We built the dataset of the six object categories by taking pictures and using photo downloaded from the Internet, especially Weixin official account platform. To maximise diversity, the images include variations in shooting angles, lighting conditions and background. We totally collected ~2,000 pictures.

During the annotation process, we used LabelMe and Robofolow to accurately label the targets and took advantage their enhancement features including rotation, stretching, sharpening and colour jitter to extend the dataset. We also added preprocessing steps to the pipeline: images were first resized to 256×256 pixels then cropped to 224×224 pixels to ensure uniform input dimensions and finally normalised. Finally, a corresponding JSON segmentation annotation was generated for each image.

During testing, we adopted a Test-Time Augmentation (TTA) strategy, including horizontal flipping. Each image was inferred twice including original one and flipped one, and the prediction probabilities were averaged. This helps reduce the error caused by viewpoint bias and enhances the model's generalisation, proven to improve accuracy gain of 1~2%.

For data loading, we defined a custom class CustomDataset which directly reads the JSON segmentation annotations for each image and replaced the background to white, effectively isolating object regions. The background removal strategy ensured the model to focus on salient object features and minimizing background noise.

III. Model Choosing and Design

In this project, we utilize a dual-model framework incorporating both ResNet50 and Vision Transformer (ViT).

$$y = F(x, \{W_i\}) + x \qquad (1)$$

ResNet50, a deep convolutional neural network, employs residual connections as seen in (1) to effectively mitigate the vanishing gradient problem commonly encountered in deep networks. It has demonstrated strong performance in image classification tasks, especially on medium to large-scale datasets. Its mature design, training stability and effective transfer learning make it become a widely recognised choice.

ViT, based on the Transformer architecture. It excels at capturing long-range dependencies within images and has emerged as a powerful alternative to traditional CNNs in visual recognition tasks. Its capacity to process large-scale datasets enables the learning of more expressive and diverse feature representations.

$$\hat{y} = \frac{1}{2}(Softmax(f_{ResNet}(x) + Softmax(f_{ViT}(x)))) \qquad (2)$$

The combination of ResNet50 and Vit is motivated by their complementary strengths: CNNs like ResNet are adept at extracting local features and Transformers such as ViT are proficient in modelling global context. For ease use, the train.py script has been configured for one-click training of both ResNet50 and ViT. Then automatically saving the results as best_model.pth and vit_model.pth respectively. During inference, the pre-trained ResNet50 and ViT models are loaded and integrated via an EnsembleModel. In the forward pass, SoftMax is applied to each model's output and their probability distributions are averaged as seen in (2) to computed to yield a more reliable prediction.

IV. Training and Validation

The dataset is split into training sets and validation sets at 8:2 ratio with images and annotations stored in the train_jpg/tain_json and val_jpg/val_json directories respectively. The CustomDataset class handles mask synthesis within its __getitem__() method, removing the background to ensuring that only the foreground object of interest in the output image. Data loading is handled by the get_dataloaders() function which allows the batch size to be configured via the --batch_size parameter. The number of output classes is automatically inferred from the training dataset to ensure consistency with the model's output layer.

$$L = -\sum_{i=1}^{C} y_i log(\hat{y}_i) \quad (3)$$

$$\eta_t = \eta_{min} + \frac{1}{2}(n_{max} - \eta_{min})(1 + \cos(\frac{T_{cur}}{T_{max}}\pi)) \quad (4)$$

Training utilizes the CrossEntropyLoss function as seen in (3) as the objective, optimized using the AdamW optimizer in combination with a Cosine Annealing learning Rate Scheduler (CosineAnnealingLR) as seen in (4) to achieve stable and smooth convergence.

Each training epoch consists two phases: training phase tran() and validation phase eval(). In the training phase, we computed the loss of each batch, backpropagated gradients and updated parameters. Int the validation phase, we disabled the gradient computation and evaluated the model's performance on the validation set. At the end of each epoch, the model's parameters are updated only if a new best validation accuracy is achieved to ensure that the best preforming model is retained.

V. Performance

On our internal validation set, the ResNet50 individual model achieved 100% accuracy and ViT model attained similar results through the same training process. After applying ensemble learning with TTA (Test-Time Augmentation), the test set accuracy consistently converged to over 95%, exceeding then individual model (approximately 92%). The ensemble helps to offset each model's weaknesses: ResNet50 is more robust in low-light or high-noise scenarios and ViT is more sensitive to fine details in small samples. The combination greatly improves generalisation. According to the TA's testing result, our model achieved an accuracy of 90%.

VI. Performance Analysis

We employed multiple approaches to optimize model performance, including using a DataLoader with num_workers=4 to achieve efficient parallel I/O loading; using the AdamW optimizer and the CosineAnnealingLR scheduler to ensure smooth convergence during training; retaining only the target region after background removal to reduce irrelevant information and boost both training and inference speed; applying cross-architecture ensembling (ResNet50 and ViT) combined with TTA (Test-Time Augmentation) using horizontal flipping to significantly improve the model's generalisation and robustness; and saving only the best-performing model parameters during training to conserve storage space and speedup loading.

To gain a deeper understanding of the model's decision-making mechanism and to evaluate its performance, we conducted the following experiments:

A. Feature Extraction

Objective: to intuitively compare the feature extraction mechanisms of ResNet50 and ViT, offering a basis for ensemble and TTA strategies. Verifying whether models are properly trained or suffering from overfitting or underfitting.

In this experiment, we first registered intermediate feature hooks in ResNet50 and forcibly extracted attention weights from ViT's MultiheadAttention module using a registered hook. Then the images are preprocesses and forwarded through the models, triggering the hooks and visualized intermediate ResNet50 features and ViT attention rollout.

To better understand the internal working of both models, we extracted intermediate feature maps from ResNet50 to ViT. ResNet50's maps show transition from edge/texture detection (shallow layers) to object-level abstraction (deeper layers). ViT, using its transformer-based architecture, captures longer-range dependencies. This helps us understand how both models extract the useful data from raw images.

Fig. 1. Selected library raw photo

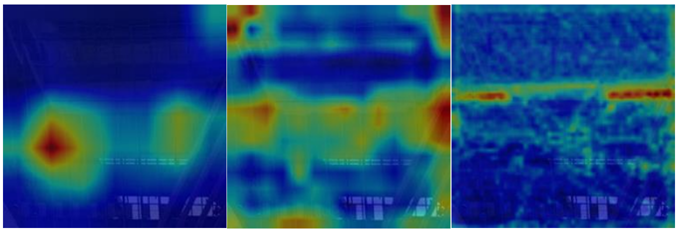

Fig. 2. Heatmaps: ResNet-library in order to Layer 1, Layer 3 and Layer 4

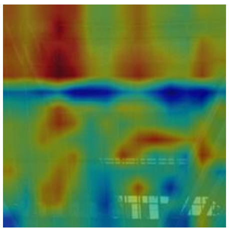

Fig. 3. Heatmaps: ViT-library

In ResNet50, Layer 1 focuses on local details like edges, Layer 3 start capturing semi-global structures and highlights object contours while suppressing background. In ViT, the attention heatmaps are generated by multiplying multiple attention layers and aligning [CLS]-to-patch attention. Red regions in the heatmaps indicate the most influential patches for classification. The heatmaps strongly overlap with the object shape, especially sharp tips, indicating effective learning. There results support the visual complementarity of ResNet50 and ViT, justifying an ensemble approach.

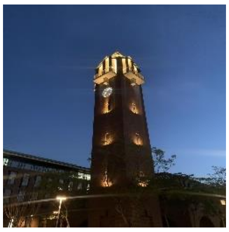

Fig. 4. Selected bell-tower raw photo

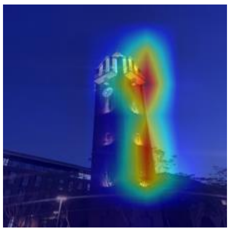

Fig. 5. Heatmaps: ResNet of library

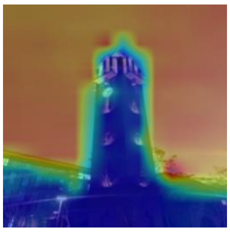

Fig. 6. Heatmaps: ViT of library

$$L^{c}_{Grad-CAM} = ReLu(\textstyle\sum_k a^{c}_{k} A^{k}) \qquad (5)$$

Final activation maps using Grad-CAM illustrate where models focus on during classification. For example, when recognizing a bell tower, both ResNet50 and ViT highlight the spire and main structure, showing they focus on discriminative regions.

Analysis: Experiment showed that ResNet50 might fail under occlusion or noise while ViT still captures object shape. And ResNet50 tends to focus only on visible textures. Therefore, the "ResNet50 + ViT Ensemble" is theoretically supported: when one fails, the other compensates.

B. *Ablation Study*

Objective: to evaluate the importance of different components, optimize the model architecture, and improve robustness.

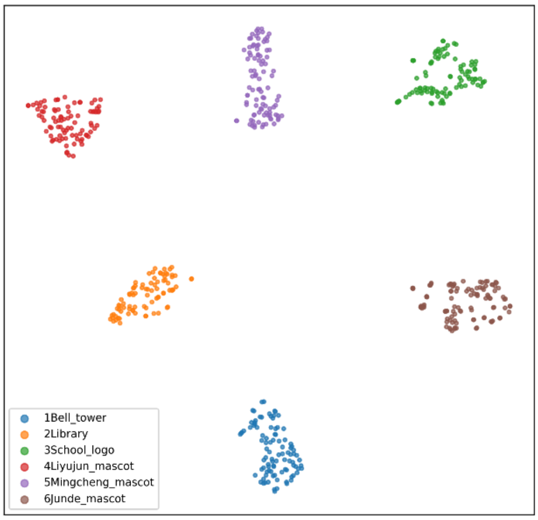

Fig. 7. t-SNE of ResNet50 Features

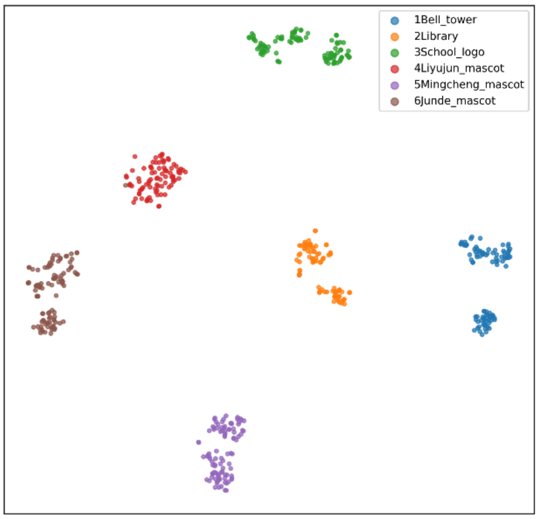

Fig. 8. t-SNE of ViT Features

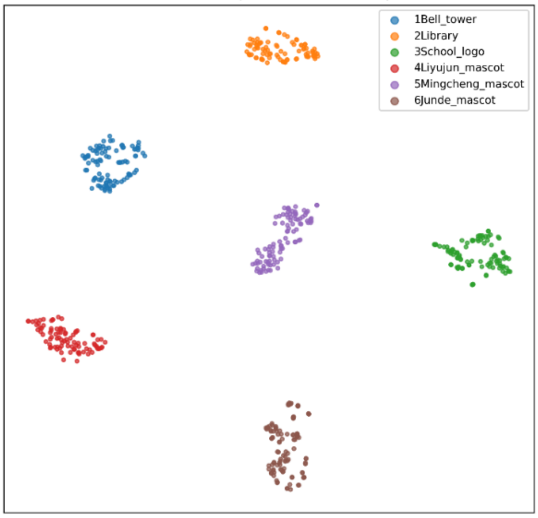

Fig. 9. t-SNE of Ensemble Features

To assess feature extraction effectiveness, we performed clustering analysis on model outputs using t-SNE. Fig. 9 show that tight clusters within each class and well-separated inter-class clusters, indicating high similarity within the same class and good discrimination across different classes. The ensemble model, combining ResNet50's locality and ViT's global view, achieves better feature separability and stronger generalisation to unseen samples.

Mode	***ResNet50 (%)***	***ViT (%)***	***Ensemble (%)***
Original	100.00	99.67	100.0
Erasing	98.33	97.33	100.00
Noise	81.33	99.33	99.33
Brightness	99.67	99.67	100.00

Fig. 10. Accuracy under adversarial conditions with ensemble performance

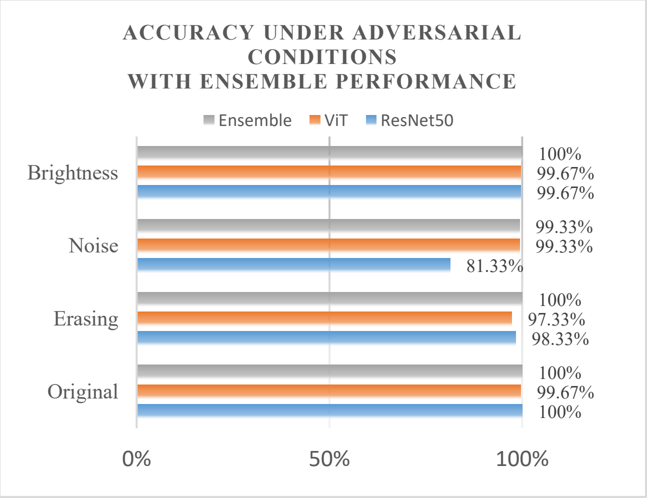

Fig. 11. Accuracy under adversarial conditions with ensemble performance

We tested models' robustness under adversarial conditions included occlusion, noise, brightness variation.

On clean images, all three models both performed well and ensemble model even achieving 100%. Under occlusion, ResNet50 remained robust; ViT's accuracy dropped more significantly, ViT was more sensitive; the ensemble achieves 100%, leveraging strengths of both. Under noise, ResNet50 dropped significantly; ViT remained stable; the ensemble matched ViT's performance, avoiding failure. Under brightness variation, all models stably; the ensemble fixed rare error, reaching 100%.

C. Test-Time Augmentation Study

Objective: evaluate the effectiveness of Test-Time Augmentation (TTA) in enhancing image classification accuracy.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N}(Softmax(f(x_i))) \qquad (6)$$

TTA refers to the technique of applying multiple different augmentation transformations to the same test image during inference. These transformed versions are fed into the model, and their predictions are aggregated (usually taking averaging as seen in (6)) to generate the final classification result.

Model predictions may vary slightly when the same image is subjected to different transformations such as cropping, flipping or scaling. TTA helps smooth out these fluctuations and improves overall prediction robustness. By applying various distortions (e.g., different scales and angles) to test images, TTA also reduces biases caused by differences in lighting, focus or composition.

In our experiment, each test image was augmented using 5 different transformations before being input into the model. The results are as follows:

Ensemble Baseline (without TTA) Accuracy: 88.33%

Ensemble with TTA (5 crops + flips) Accuracy: 90.00%

TTA resulted in an approximate 2% improvement in accuracy. Given the limited size of the test dataset, we expect that the improvement would be more substantial in a larger-scale evaluation.

VII. Innovation

Removing background for each image. Therefore, model can focus on targets and reduce background interference. Integrating Convolutional Neural Network (ResNet) and Transformers (ViT) for joint training and inference, leveraging the complementary strength of both architectures. Applying Test-Time Augmentation (TTA) via horizontal flipping during inference to enhance model robustness through multi-view reasoning. Developing one-click training pipeline to simplifying workflow and improving operational efficiency. Optimizing data loading and model architecture for faster training and inference processes.

VIII. Individual Workload of Students

A. ***

Conducted preliminary literature review; Developed data preprocessing scripts (including but not limited to data augmentation and background removal using masks) Built, selected and trained models; Designed and conducted experiments; Making 9 charts for Report; Assisted in refining the report and presentation slides; Took charge of the presentation and Q&A session.

B. *** (repository maintainer)

Coded a Web Scrap for Sogou; Took, collected and annotated over 1600 photos; Translated, revised and added content to Report; Made 1 chart and 6 formulas for Report; Proofread and added content to PowerPoint; Edited and added content to speech script.

C. ***

Took, collected and annotated approximately 300 images; Drafted the initial Chinese version of the report, created presentation slides and the speech script; Making 1 chart for Report; Presented the content using the slides.

IX. Discussion and Conclusion

We improved the accuracy and robustness of image classification through innovative data preprocessing, model fusion and augmentation strategies. Removing the image background to reduce noise. Used cross-architecture integration effectively combined the strengths of different models. The Test-Time Augmentation (TTA) strategy further improves the model's generalisation capability.

By visualising final activation maps, intermediate decision-making processes, classification results and two experiments-feature extraction and ablation study, we deepened our understanding of the model's decision logic and we evaluated its performance.

We are looking forward to exploring more diverse data augmentation techniques, deeper model fusion approaches and adopting more efficient training techniques to further improve performance.